Hybrid Machine Learning and Ensemble Data Assimilation

Wei Pan

Co-authors and contributors: Massimo Bonavita, Elías Hólm, Marcin Chrust, Alban Farchi, Marc Bocquet, Patrick Laloyaux, Ivo Pasmans, Sebastien Massart, Boštjan Melinc, Žiga Zaplotnik, Simon Lang, Ewan Pinnington

5th ECMWF-ESA ML Workshop, Bologna, 13-17 April 2026

ECMWF

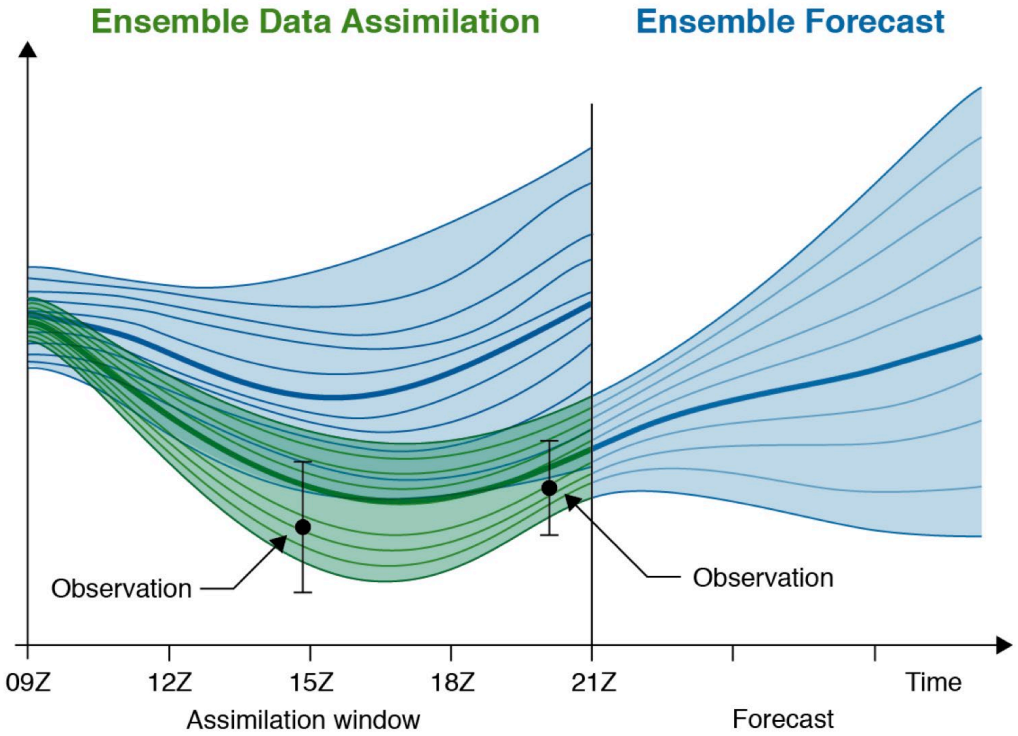
Ensemble 4D-Var

4DVar in Bayes form:

$$p(\mathbf{x} \mid \mathbf{y}_{\text{obs}}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)\right) p(\mathbf{y}_{\text{obs}} \mid \mathbf{x})$$

Variational DA produces *maximum a posterior* (MAP) estimates

$$\mathbf{x}^a = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}).$$



Ensemble DA (EDA) produces a collection of many perturbed MAP optimisations

$$\mathbf{x}^{a,(i)} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}^{(i)}, \mathbf{x}^{b,(i)}, \boldsymbol{\eta}^{(i)}), \quad i = 1, \ldots, N.$$

$\boldsymbol{N = 50}$ at ECMWF

As well as providing ICs for ensemble forecasting, the EDA provides **flow-dependent estimates of analysis and background error uncertainty** (*Isaksen et al* (2010)),

$$\mathbf{B}(t) \approx \frac{1}{N-1}\sum_{i=1}^{N}[\mathbf{x}^{b,(i)}(t) - \bar{\mathbf{x}}^b(t)]\,[\mathbf{x}^{b,(i)}(t) - \bar{\mathbf{x}}^b(t)]^T.$$

Hybrid 4DVar-ML EDA

Aim – Develop an EDA emulator and integrate it into the operational EDA workflow.

Two complementary directions:

1. EDA statistics
    - Emulate full-EDA statistics using fewer members.

$$\{\mathbf{x}^{b,1}, \mathbf{x}^{b,2}, \ldots, \mathbf{x}^{b,N}\} \longrightarrow \mathbf{B}$$

approximates

$$\{\mathbf{x}^{b,1}, \mathbf{x}^{b,2}, \ldots, \mathbf{x}^{b,M}\} \longrightarrow \mathrm{ML} \longrightarrow \tilde{\mathbf{B}}, \quad M < N$$

2. EDA members generation
    - Generate EDA members that effectively mimic the full 4DVar EDA.
        - Full 4DVar EDA ensemble $\{\mathbf{x}^{a,1}, \mathbf{x}^{a,2}, \ldots, \mathbf{x}^{a,N}\}$
        - Hybrid 4DVar and ML EDA ensemble $\{\mathbf{x}^{a,1}, \mathbf{x}^{a,2}, \ldots, \mathbf{x}^{a,M}\} \cup \{\tilde{\mathbf{x}}^{a,j_1}, \tilde{\mathbf{x}}^{a,j_2}, \ldots, \tilde{\mathbf{x}}^{a,j_{N-M}}\}$

ML emulated members

Hybrid 4DVar-ML EDA

Current status

1. EDA statistics
    - Developed and evaluated a prototype emulator for operational 50-EDA sampled stDev $\Sigma(t)$ (**es**) , using input ensemble size of **n=5** (*ECMWF Tech Memo 936*).

2. EDA members generation
    - Developed and validated a ML methodology for EDA analysis perturbations using the Lorenz-96 low-order model as a testbed.
    - A scaled prototype of the method at higher resolution is under development (using Anemoi) and testing.
        - Currently working on integrating the model into our EDA workflow for testing, evaluation and improvements.

EDA statistics

EDA variance emulator

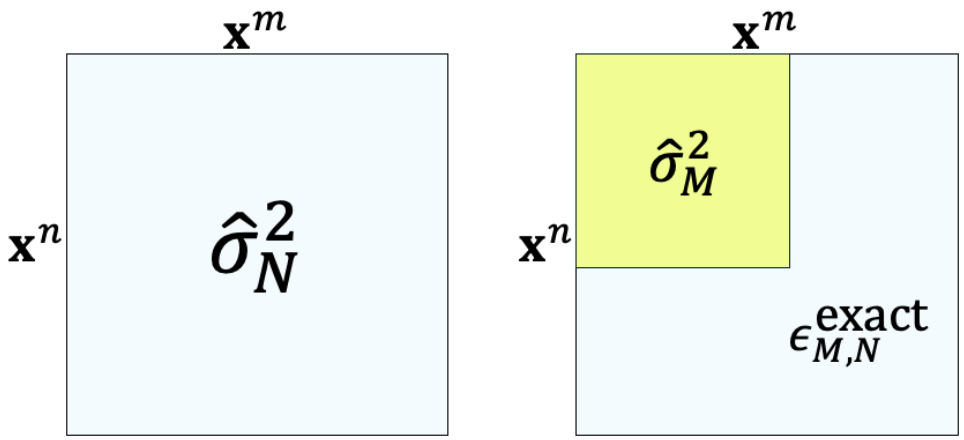
See ECMWF Tech Memo 936 for details; doi: 10.21957/0b7e4d4426

- The variance **Σ** part of **B** is just the pointwise unbiased sample variance

$$\hat{\sigma}_N^2 := \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}^{b,(n)} - \bar{\mathbf{x}}^b)^{\otimes 2} = \frac{1}{2N(N-1)}\sum_{n,m=1}^{N}(\mathbf{x}^{b,(n)} - \mathbf{x}^{b,(m)})^{\otimes 2}$$

- Fix a small ensemble estimate, M < N

where



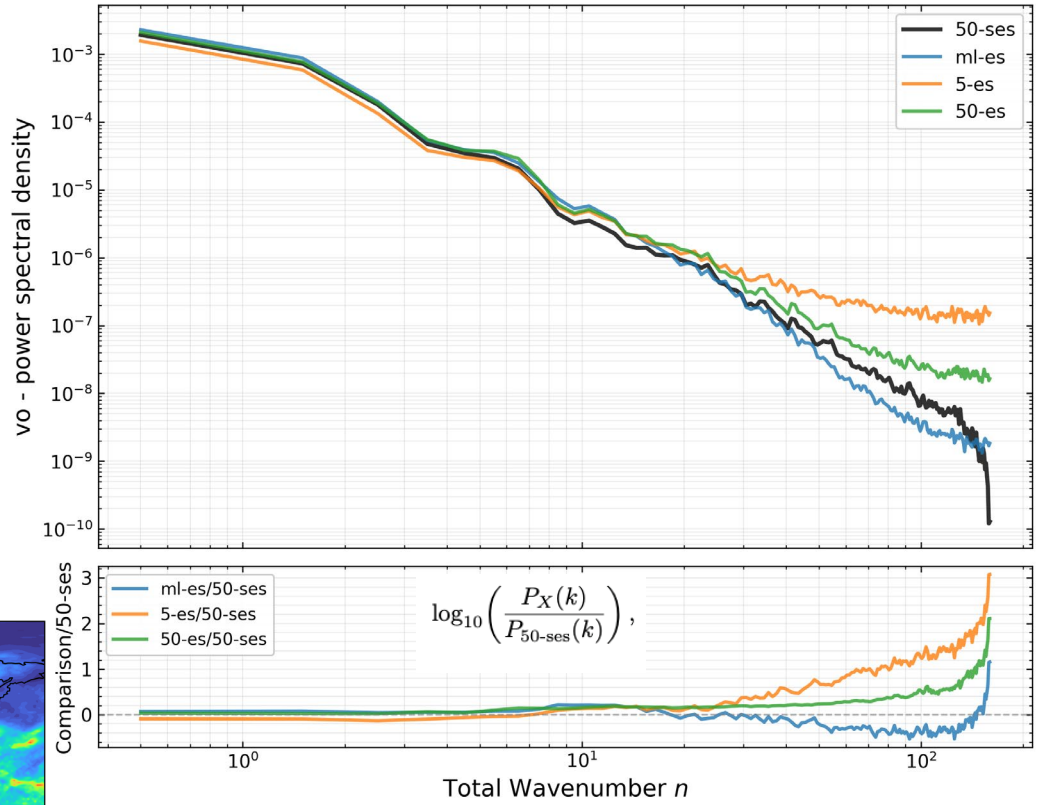
- The proposed model takes as input $\hat{\sigma}_M$. Its output is a probability distribution on $\hat{\sigma}_N$

$$\hat{\sigma}_M \mapsto \mathbb{P}_\theta(\sigma_N \mid \hat{\sigma}_M).$$

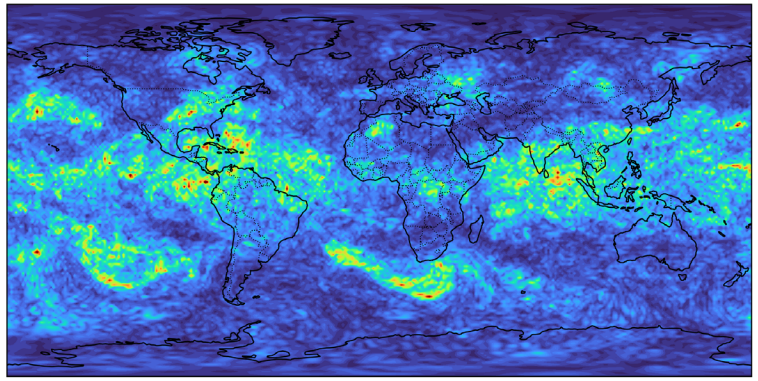
EDA variance emulator

$$\tilde{\sigma} = g_{\theta_2}\left(f_{\theta_1}(\hat{\sigma}_M), \epsilon\right), \quad \epsilon \sim q_{\theta_3}\left(\cdot \mid \hat{\sigma}_M\right)$$

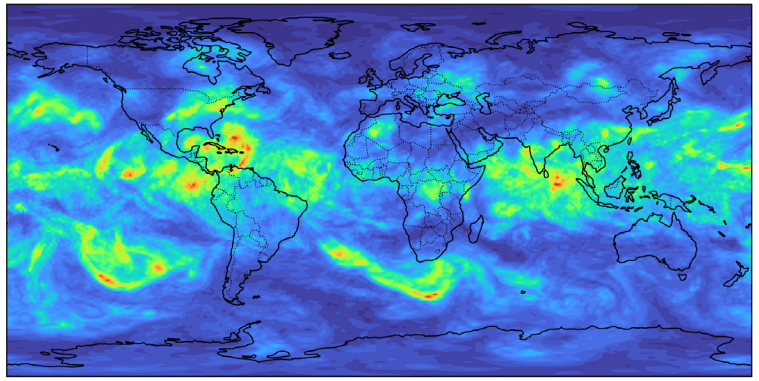
decoder — encoder — latent error distribution



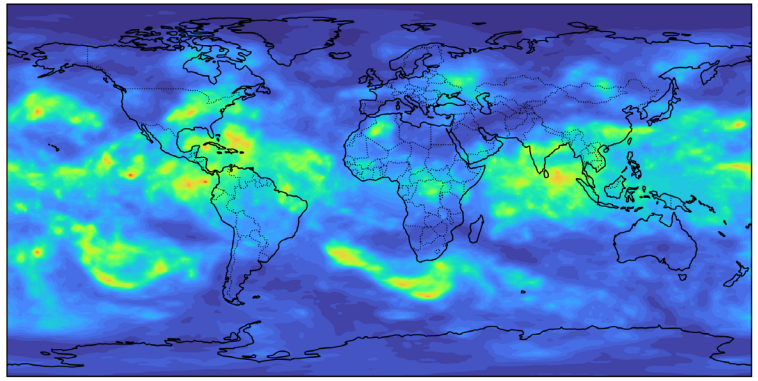
5-es



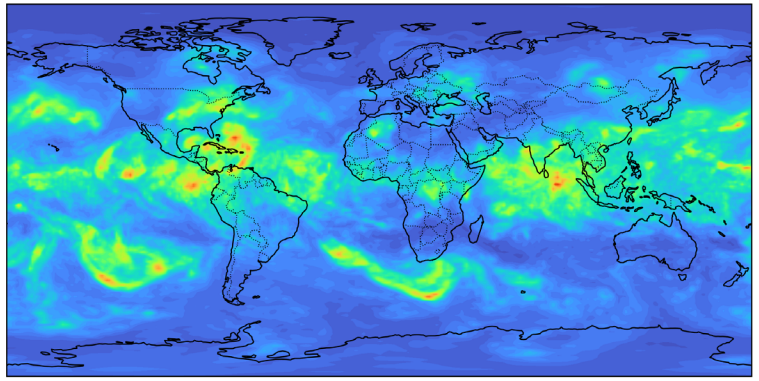
50-es



ml-es



50-ses



M=5, N=50
N80 grid, 137 model levels
6 variables: (vo, div, t, lnsp, q, o3)

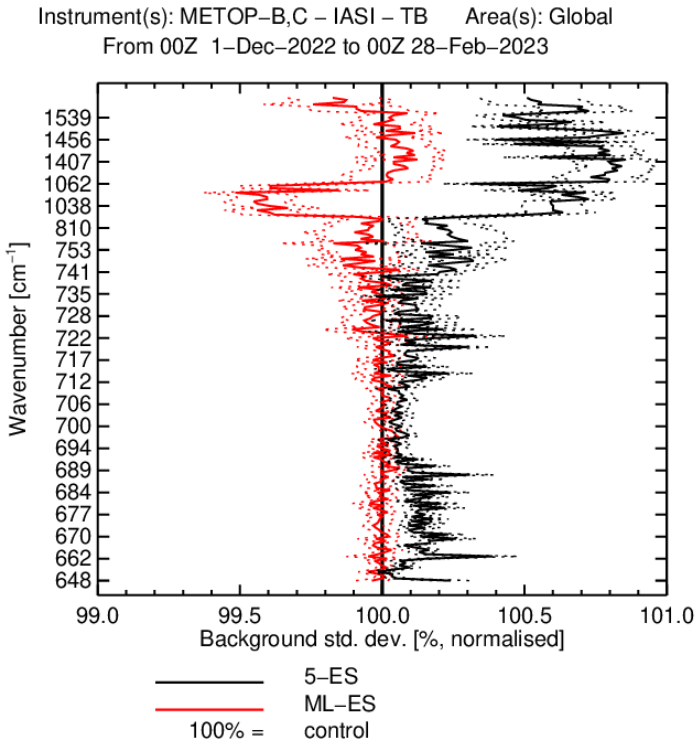
The training dataset spans
01/01/2023 – 31/12/2023.

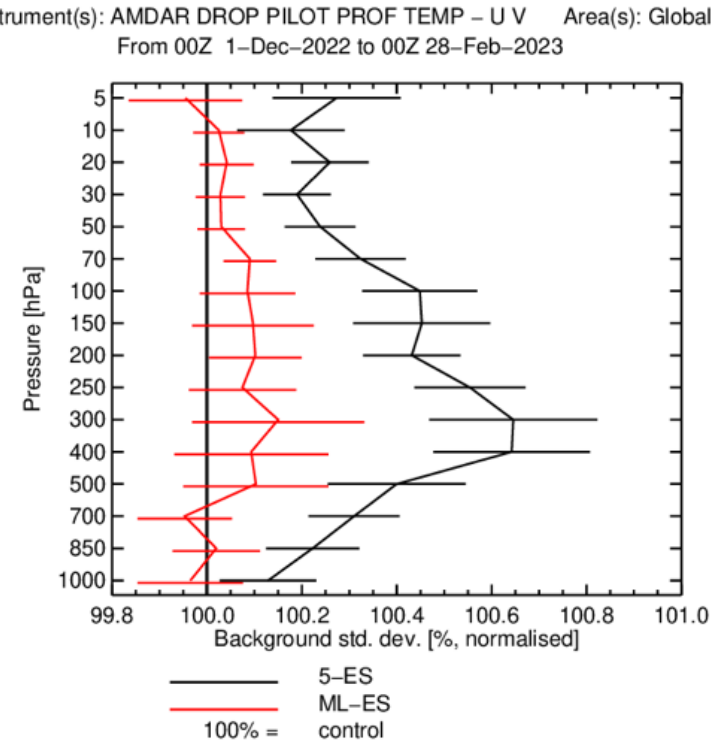
Each model has
< 200,000 learnable parameters

e.g. Vorticity es, ml74 (~200hPa)
21:00 UTC 01 June 2022

Normalised observation statistics (obstats)

Satellite instruments





Conventional instruments

- Perfect replication is ***not possible*** due to information theoretic lower bounds

$$\mathbb{E}\left[(\bar{\sigma}-\hat{\sigma}_N)^2_{i,(\lambda,\phi)}\right] \geq \mathbb{E}\left[(\mathbb{E}(\hat{\sigma}_N|\mathscr{F}_M)-\hat{\sigma}_N)^2_{i,(\lambda,\phi)}\right],$$

any estimator from M samples

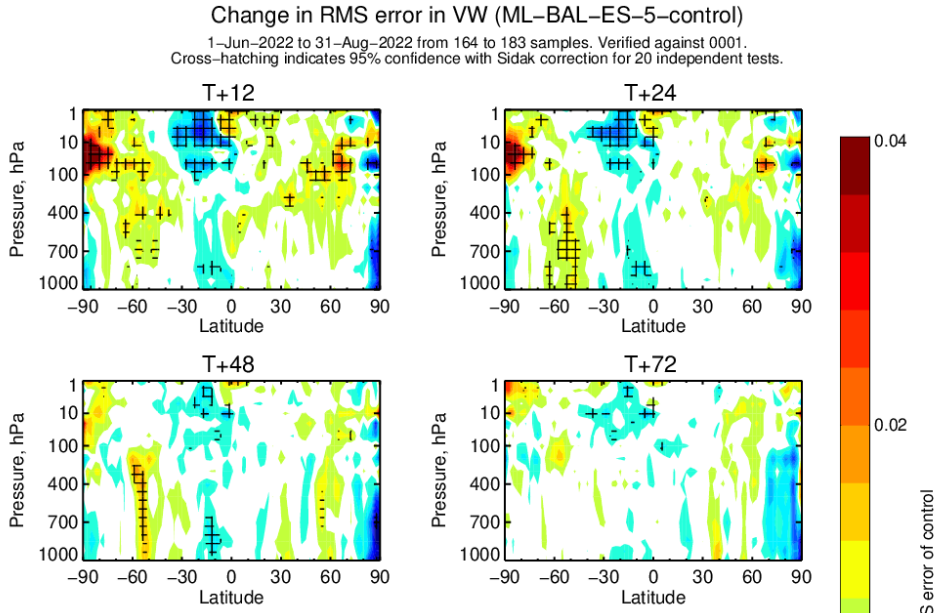
Optimal with respect to mean square error

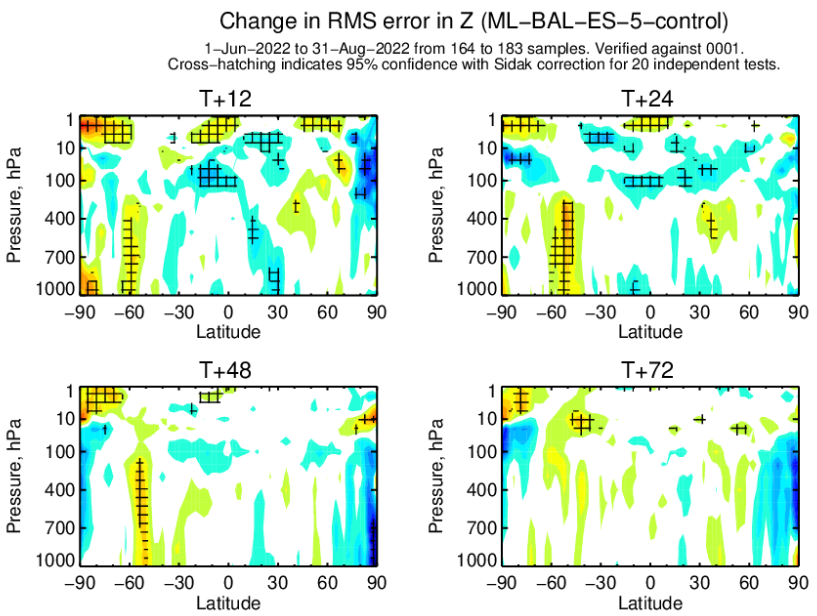
- 4DVar optimisation is **continuous** with respect to **B**, i.e. a small change in **B** means a small change in the resulting analysis increment.

Normalised change in RMS forecast error

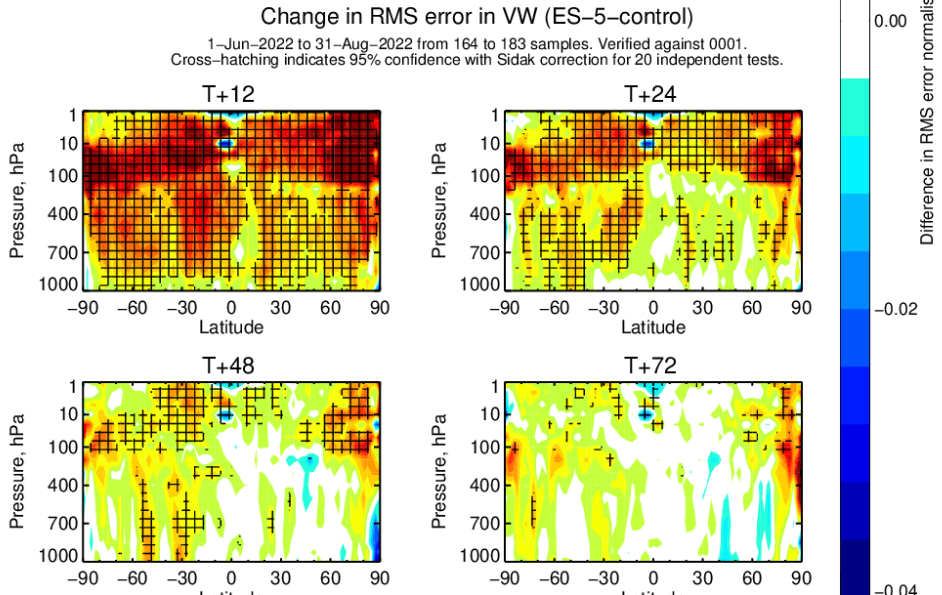
ML-es from 5 members

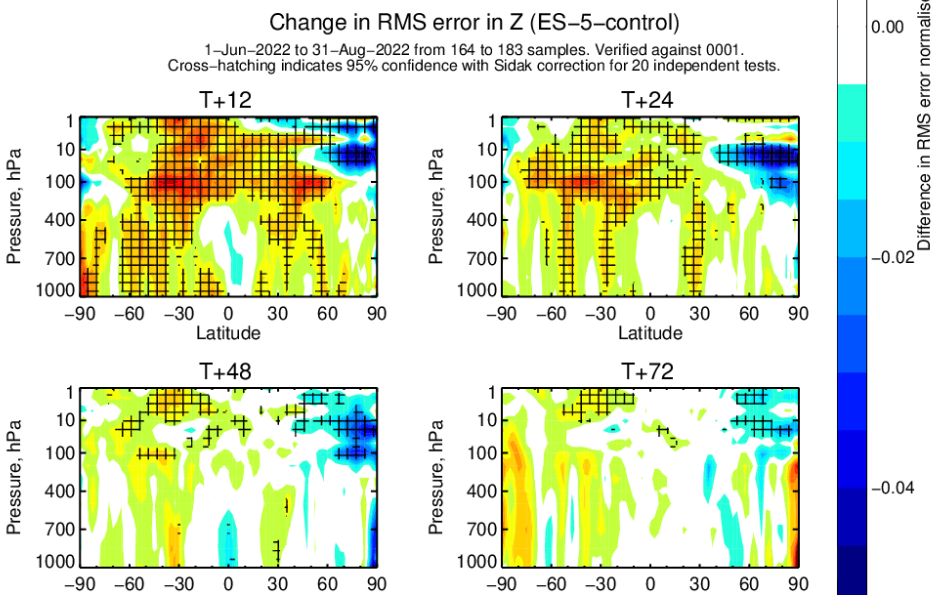
vs 50-es





5-es vs 50-es





EDA members generation

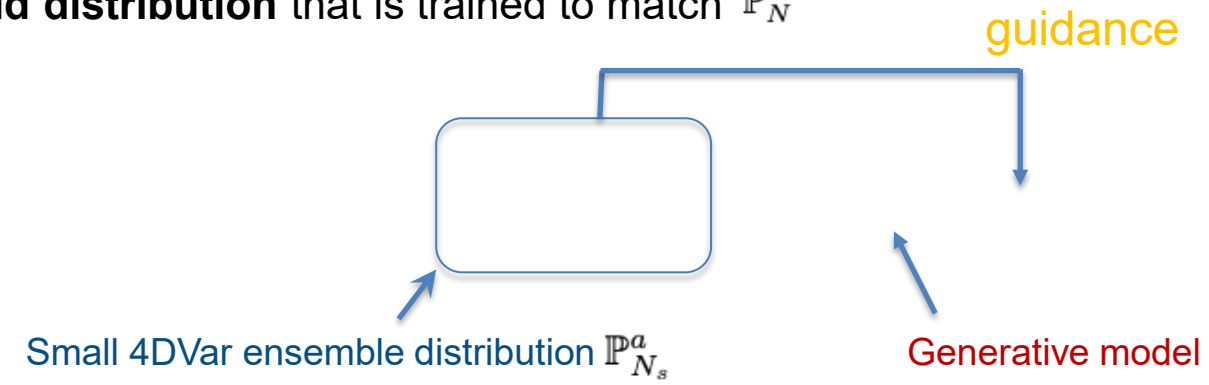
EDA Emulation

The EDA system provides empirical background and analysis distributions

$$\mathbb{P}_N^b := \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{x}^{b,(i)}}, \qquad \mathbb{P}_N^a := \frac{1}{N}\sum_{i=1}^{N}\delta_{\phi(\mathbf{x}^{b,(i)}|\mathbf{y}_{\text{obs}})}$$

where ϕ denotes the 4DVar transformation applied to the background states.

We wish to learn a **hybrid distribution** that is trained to match $\mathbb{P}_N^a$

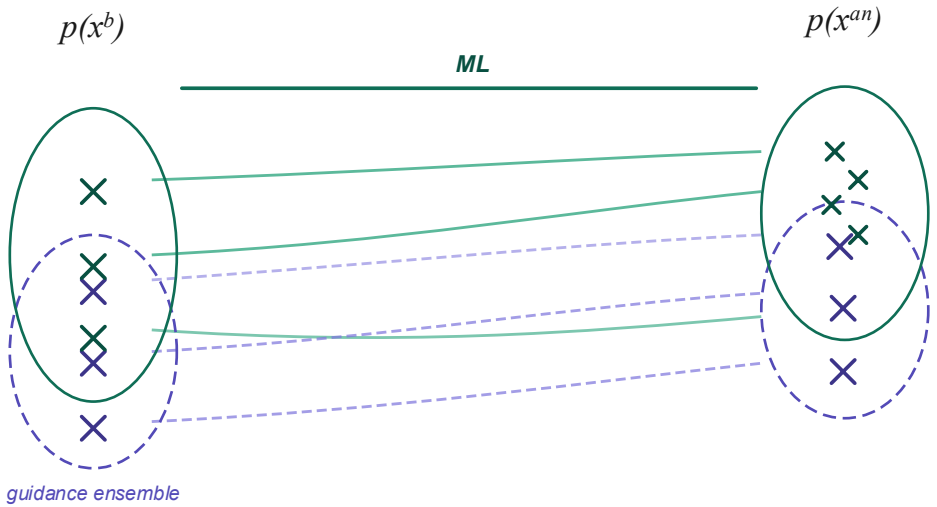


We propose a **stochastic flow** model (stochastic transport). Stochasticity enables cheap ensemble enlargement.

$$\phi_t^{(j)} = u_\theta(\phi_t^{(j)}, \mathbb{P}_t, t \mid \mathbb{P}_{N_s}^a)dt + \sigma_\theta(\phi_t^{(j)}, \mathbb{P}_t, t \mid \mathbb{P}_{N_s}^a) \circ dW_t, \quad \phi_0^{(j)} \in \left\{\mathbf{x}^{b,(k_j)}\right\}_{j=1}^{N-N_s}$$

EDA Emulation

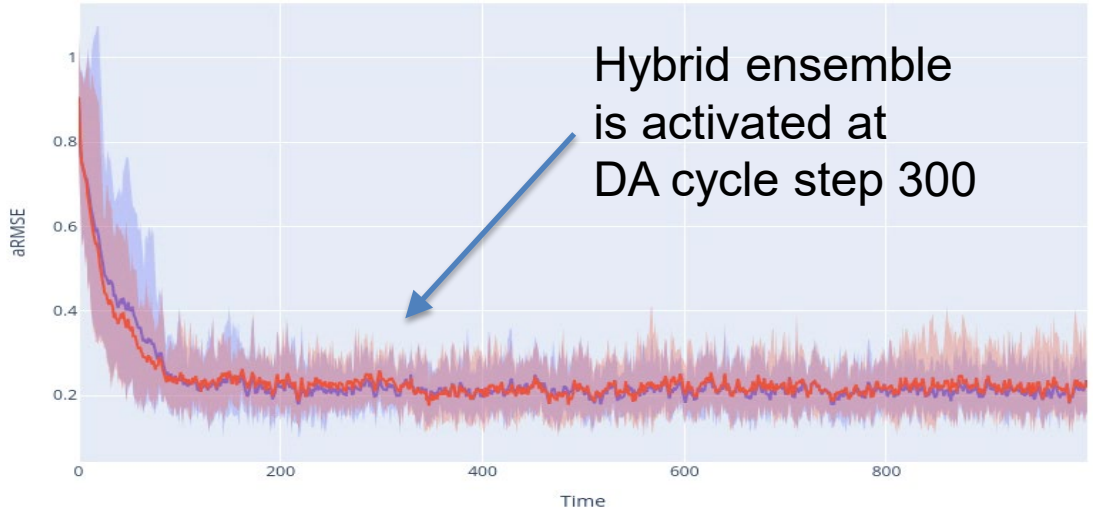
Intuitive picture – the small ensemble of 4DVar analyses guides the generative model, so that observations are assimilated indirectly.



$$\tilde{\mathbf{x}}^{\mathrm{an}} \sim \text{“}\mathbb{P}_\theta(\mathbf{x} \mid f(\mathbf{y}_{\mathrm{obs}}))\text{”}$$

guidance data contain obs info

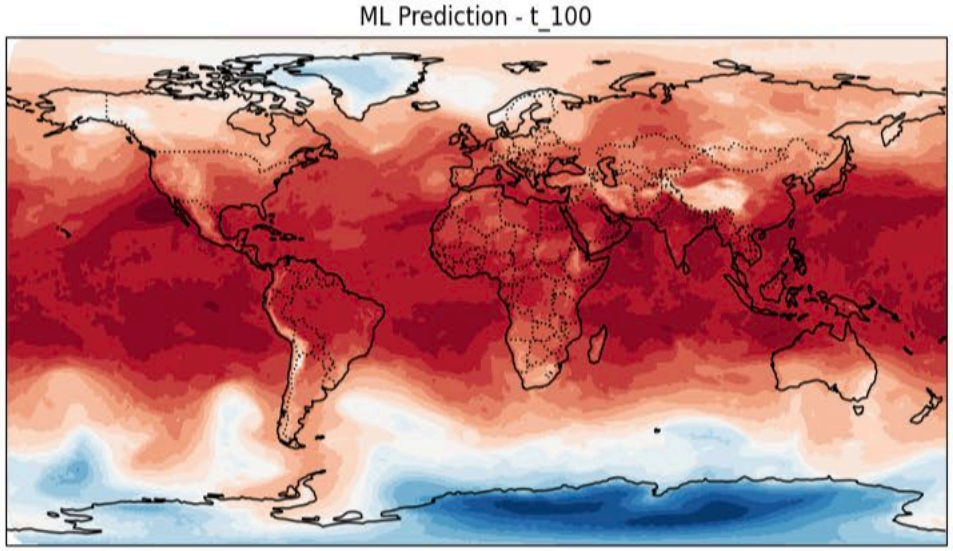
- **Lorenz96 ETKF** (from Alban Farchi)
  - dim=40, N=20 (theoretical minimum ens. size 14)
  - Inflation only, no localisation
- Hybrid ensemble $N_s = 10$ and 10 emulated.
- Neutral RMSE scores – Hybrid EDA vs Full EDA



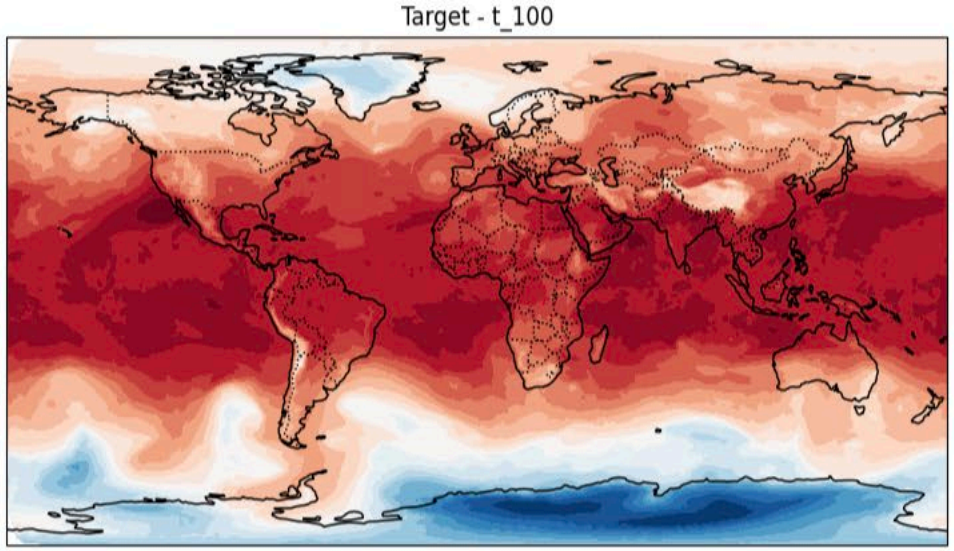
ECMWF

EDA Emulation: First Results

an[i-1] → an[i]



emulated member

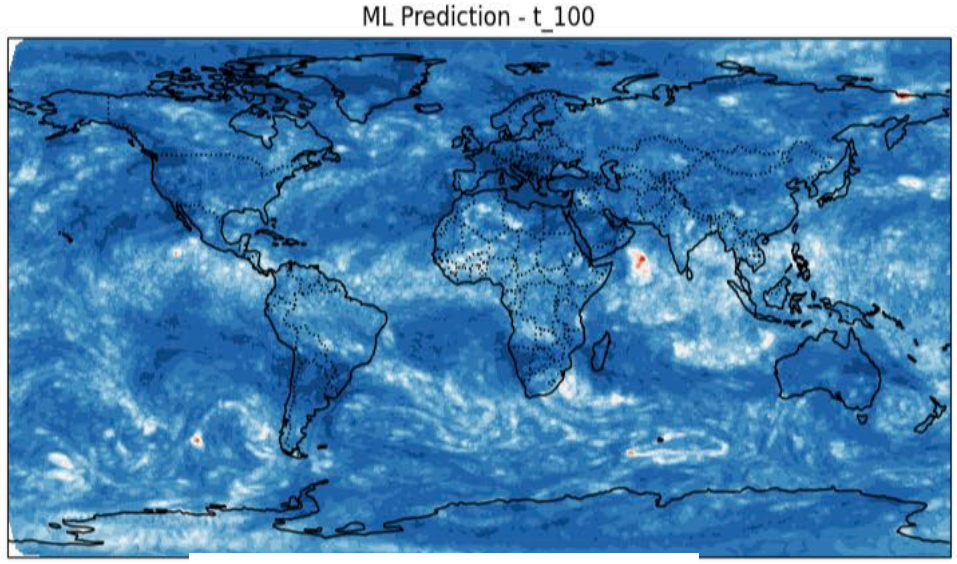


EDA member

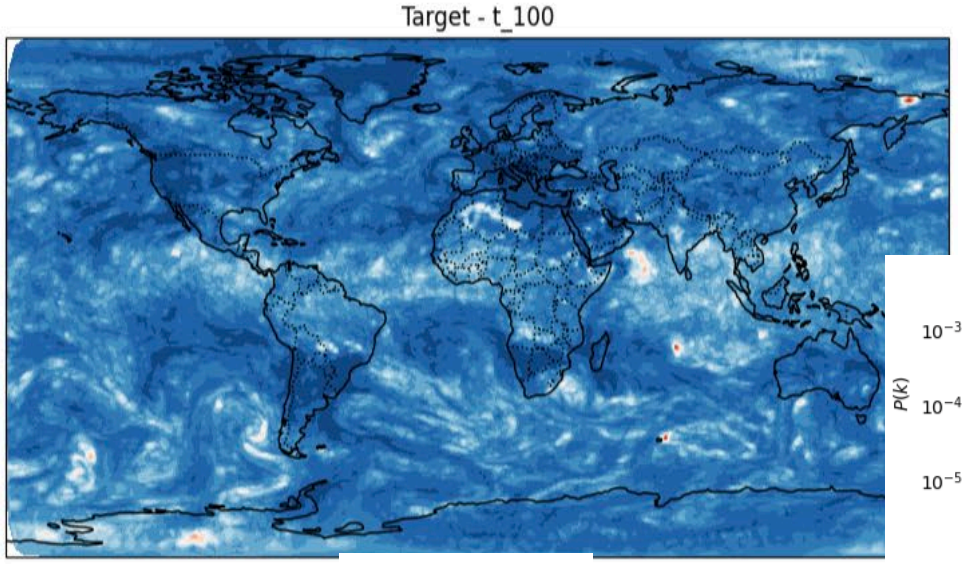
Model trained using 4 ensemble members

ECMWF OD Dataset
2024.07.01 – 2025.06.30

(t, vo, div, q, o3, lnsp) full model levels, N128 grid.

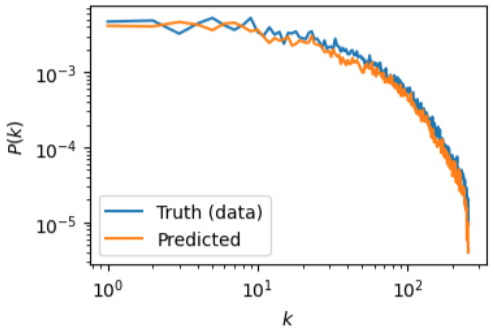


es (14 EDA + 36 emulated)



es (50 EDA)

Resolution is consistent with 50R1 an increments.



Thank you for your attention.